Politecnico di Torino

Université Paris Cité

Master of Science in Physics of Complex Systems

Final project

Allostery and hydration dynamics: a molecular dynamics study of dihydrofolate reductase

École normale supérieure

Department of Chemistry, UMR ENS-CNRS-UPMC 8640,
École Normale Supérieure, 24 rue Lhomond, 75005 Paris, France

Supervisor:
Prof. Andrea Antonio Gamba, Politecnico di Torino
Co-supervisors:
Prof. Damien Laage, ENS
Prof. Olivier Rivoire, Collège de France
Prof. Clément Nizak, Collège de France
Prof. Guillaume Stirnemann, IBPC

Author:
Salvatore Di Marco

Academic year 2021–2022

ENS | PSL

Collège de France 1530

IBPC

Abstract

Allostery is a biological phenomenon which is displayed in many different proteins, and consists in any kind of coupling between the active site of the protein and a distant site. There is no unique way to identify allosteric sites, even with experimental analyses. This work, by means of Molecular Dynamics simulations, will focus on dihydrofolate reductase (DHFR). We attempt to identify a link between two different classes of amino acids. The first are called sectors: they were obtained by evolutionary data and have been shown to have strong superpositions to allosteric sites. The second ones display a coupling to the catalytic activity of the protein, and were obtained in this work by means of molecular dynamics simulations. We show that the couplings in the second class of amino acids are quite small, but we obtain that coupled sites are connected, in a network-like way, to sectors.
Moreover, we analysed, by means of simulations, the first hydration shell of DHFR, in order to verify if sectors show peculiar hydration properties. We show that there exists a strong connection between amino acids around which the reorientational times of water are largest, and sectors. If the results can be generalized to other proteins, it would be possible to make predictions of allosteric sites only by analyzing the protein hydration shells by means of simulations.

Contents

1 Background

1.1 Allostery

Enzymes are proteins which can catalyze reactions by reducing their activation energy. They are composed of an active site, where they react with molecules specific to the enzyme itself, called substrate. Moreover, enzymes can sometimes manifest the so-called allosteric regulation (Bowerman and Wereszczynski 2016; Guo and Zhou 2016), meaning that they can bind with molecules different from the substrate, and not in the active site, modifying the catalytic rate or the conformation of the protein itself.
The sites where this interaction takes place are called allosteric sites.
Allosteric regulation plays an important role in many different biological processes, and it is present in many different proteins. Haemoglobin is an example of an allosteric protein (Ahmed, Ghatge, and Safo 2020). It is a protein which is composed of four subunits, and the interaction of one subunit with oxygen leads to conformational changes for the remaining three.
In general, there therefore must be some information exchange between the active site and the allosteric one, but the distance over which the information travels can be quite large (e.g. 20Å): a mediated long-range interaction seems to be taking place.
This long-range interaction is puzzling for the physicist, since the molecular mechanism through which the information spreads is not yet clear (Wodak et al. 2019).
More generally, the definition of allostery is broad. It is not just related to molecules binding to a specific site, but to any kind of long-range perturbation which will influence the catalytic or conformational properties of the protein.
In this work, we will indeed interpret allostery as any kind of coupling between one site and the active site, given that these two sites are not close to each other.
This second kind of allostery is defined as latent. DHFR, the protein which will be analysed, is indeed not allosteric in the conventional view, but it experimentally shows allosteric properties (Reynolds, McLaughlin, and Ranganathan 2011). Understanding allostery would facilitate the synthesis of drugs (Grover 2013) which can target specific sites across the proteins, directly influencing their catalytic activity. Moreover, it would be possible to engineer special

allosteric proteins which could act as bio-sensors (Villaverde 2003).

Models As already mentioned, the exact mechanism which allows the spread of the allosteric information is not yet completely clear.
Many models have been developed in order to characterize some aspects of allosteric transitions, like Monod-Wyman-Changeux (MWC) model (Changeux 2012).
We may say that two different classes of models have been developed: the first class attempts to characterize allosteric transitions, without going into molecular detail, by looking at its conformational impact. Models belonging to the second class aim to identify allosteric sites and/or networks over which the interaction propagates.
The MWC model, belonging to the first class, describes the cooperativity of subunits present in the protein and it is a phenomenological model of allosteric transitions. It assumes that the protein can be into two or multiple states, and the population of these states can be influenced by the binding of a ligand: each state has a different propensity of binding with it.
The ligand therefore modifies the equilibrium population. By inducing a modification of the free energy of these states it can also induce a conformational change in the protein.
MWC has been regarded as an Ising model in the context of biology (Garcia et al. 2011), due to its simplicity and "unreasonable effectiveness". Despite its effectiveness in explaining allosteric transitions for many different classes of proteins (under certain hypotheses), it has to be stressed that MWC model is only phenomenological. It does not explain the underlying mechanisms which lead to the transmission of the information from the allosteric site to the active site. MWC model also fails to explain allostery for proteins which are less structured and display a strong disorder.
Models which view proteins in a more statistical way have been developed: they explain MWC model and other phenomenological models within themselves and provide different explanations of allostery (Motlagh et al. 2014; Swain and Gierasch 2006). These models consist in assuming that proteins always shift between conformations, according to their free energy landscape: the action of the ligand modifies the energy and can then induce a modification of the minima of the energy, in their value or in their position, so that proteins may equilibrate into these other configurations.

There are many models which belong to the second class (Feher et al. 2014), for instance topology analyses - analyses produced by means of molecular dynamics - and elastic network models.

Computational techniques Molecular dynamics (MD) is one of the most used techniques for trying to understand the microscopic mechanisms of many biological samples.
From these simulations it is possible, for instance, to extract allosteric sites (Bowerman and Wereszczynski 2016), and atomistic properties of allostery (Gomez et al. 2022). There are different properties through which these sites can be extracted, since, as we have already mentioned, the concept of allostery is quite general and we can therefore look at different dynamical properties of the protein.
An example of such models are Elastic Network Models, (Togashi and Flechsig 2018), which consist of a coarse-grained model of a protein, where each amino acid is connected to other ones through harmonic interactions. The perturbation of a site would then propagate, vibrationally, towards the amino acids constituting the active site.
These models are quite effective in predicting allosteric sites and networks. Nevertheless, by reducing a protein into a network of springs, they simplify it considerably and it is not possible to really understand the propagation of the allosteric signal.

Goals The purpose of this work will not be to unify the two classes of models, but we will rather analyze correlation of fluctuations of some key observables in the protein, to spot allosteric sites, by means of Molecular Dynamics simulations.
Moreover, the previously mentioned models completely neglect the influence of water on the protein, which has proven to be important for different phenomena (Royer et al. 1996; Leitner, Hyeon, and Reid 2020).
The goal of this work will then be to verify if there exists a link between allostery and peculiar hydration properties in our selected protein: DHFR.

1.2 DHFR

The protein which will be analyzed in this work is Escherichia coli dihydrofolate reductase (DHFR), shown in Fig. (1). It is an enzyme which reduces dihydrofolic acid (DHF) to tetrahydrofolic acid.
Inside the active site, a cofactor, Nicotinamide adenine dinucleotide phosphate (NADPH), interacts with DHF directly and allows the transfer of the hydride related to the mentioned reduction.
The reason why it has been chosen is the fact that there are many studies available for this protein: some of these are molecular dynamics studies (Maffucci, Laage, Stirnemann, et al. 2020; Boekelheide, Salomón-Ferrer, and Miller 2011), and some other are statistical/experimental (Reynolds, McLaughlin, and Ranganathan 2011; McCormick et al. 2021).
We will use some of these papers for comparison or we will employ some results which were obtained in them, as we will see in later sections.
The protein shows two main conformations, which are called *open* and *closed*. The main difference between these structures is the behaviour of a loop (called Met20 loop). In order to quantify these differences, the distance between the Cα[1] of the amino acids 18Asn and 45His is used, shown in Fig. (1). As can be seen from Fig. (2), the distance is higher in the open configuration and large deviations can be observed.
Since a short distance (less than 8Å) is observed in the closed configuration, water is less likely to get inside the active site, and the reaction between NADPH and DHF can take place.
The Met20 loop has indeed a strong catalytic importance, because it can change the electrostatic environment around the ligands (due to higher or smaller amounts of water molecules). A higher distance of the Met20 loop will then increase the free energy barrier of the reaction, and this is why the closed configuration is the one which shows higher catalytic activity (Maffucci, Laage, Sterpone, et al. 2020).
Higher Met20 loop distances also translate into a larger distance between the two carbon atoms (later called CC distance) involved in the reaction between NADPH and DHF, as it can be seen from Fig. (3), thereby directly changing the reaction free energy barrier.

[1] A Cα atom in a biomolecule is the first carbon atom connected to a functional group. The latter is an atom or a group of atoms which causes the peculiar chemical properties of the biomolecule.

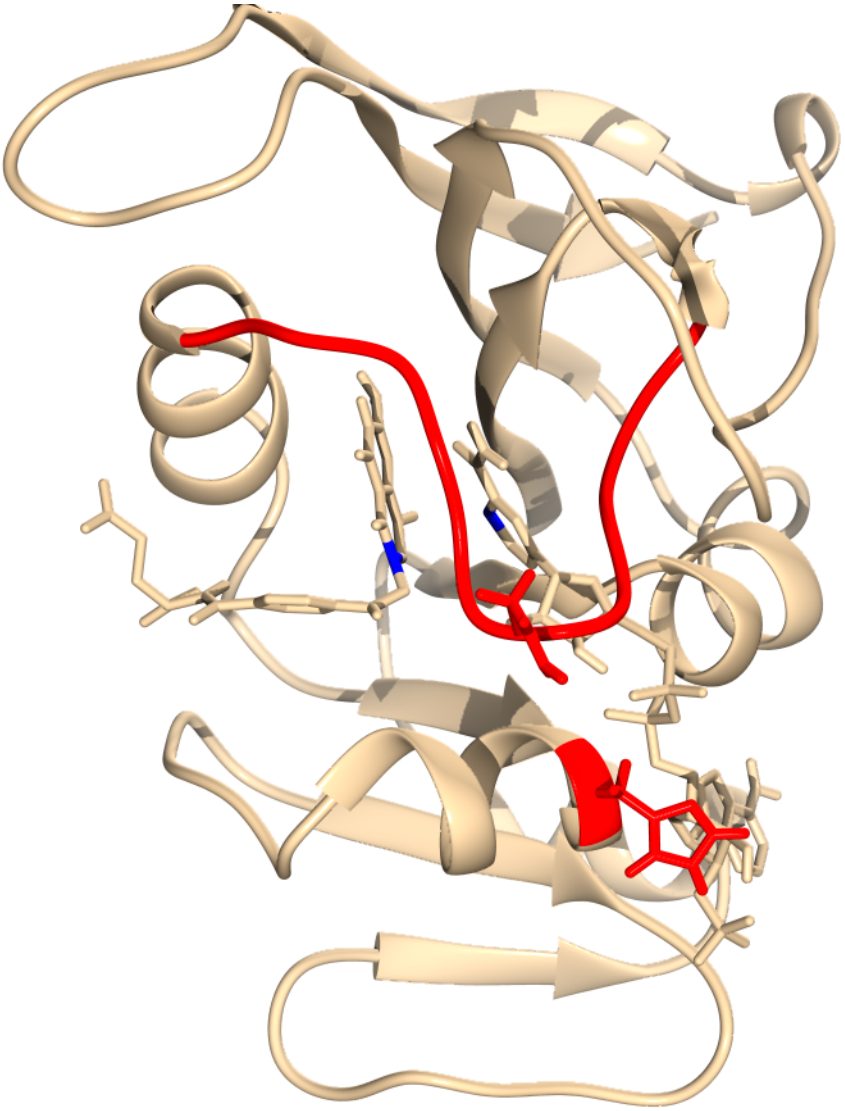


Figure 1: Closed configuration of DHFR with NADPH and DHF. The red parts highlight both the Met20 loop and 45His. Atoms of the protein are not shown, except for 18Asn and 45His, through which we arbitrarily define the Met20 loop distance. In blue, the carbon atoms relevant for the catalytic reaction, for which we compute the CC distance.

CC distance consequently has a quite strong molecular importance, because it is a direct indicator of the catalytic activity of the protein. Indeed, if the distance between the two carbon atoms are small, we expect a small free energy barrier for the reaction.

This is the reason why we will later compute correlations between the average oscillations of the Cα of the amino acids and the oscillations of this distance. A high correlation between these two quantities would imply that perturba-

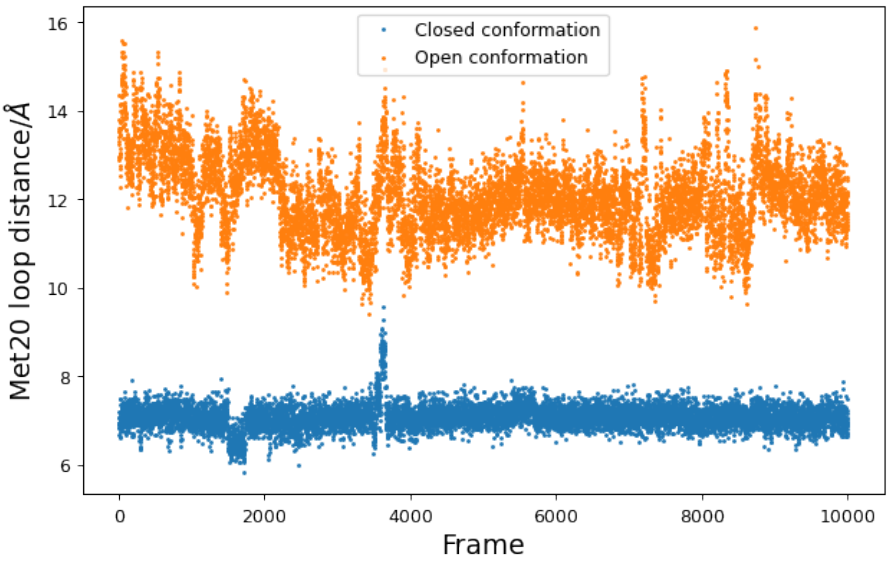


Figure 2: Distance between the Cα atoms of 18Asn and 45His, in the closed and open configurations.

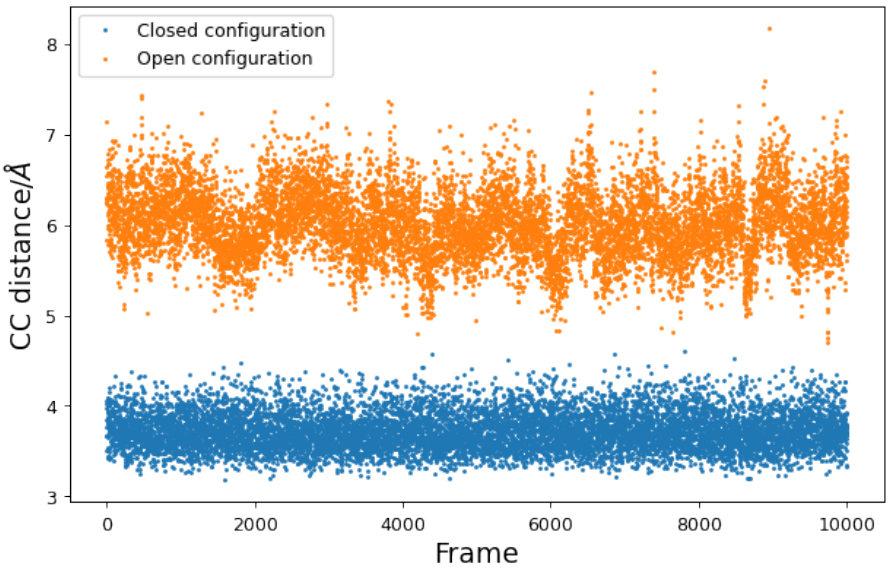


Figure 3: Distance between the carbon atoms relevant to the reaction, one from NADPH and the other from DHF, in the closed and open configurations.

tions to a given amino acid would translate into a modification of the catalytic activity of DHFR. We would therefore find allosteric sites in the protein.
Only the closed configuration will be analyzed for correlation analyses related to the active site, since the open configuration shows strong fluctuations. It is moreover not possible to identify a single equilibrated value of the CC dis-

tance in the case of the open configuration, since even in a 100ns trajectory, the value around which the CC distance oscillates changes. There are indeed at least two plateaus where the quantity is equilibrated, only to change after a lot of nanoseconds.
Since it was computationally intensive to sample all of these configurations, and to verify if there were any others, the observables we extracted for the open configuration are not expected to be physically meaningful.

1.3 Sectors and allostery

There are different methods to understand the relationship between function and structure in proteins. One of them is Statistical Coupling Analysis (SCA). Through this method (Lockless and Ranganathan 1999; Halabi et al. 2009), it is possible to find amino acids which co-evolve in a protein.
The basis of the algorithm is the assumption that proteins are always subject to random mutations, but these can happen under the constraint that the function remains largely unvaried.
A single protein can indeed be present in different species, but there will be some random variations in the chain of amino acids. Therefore, by analyzing a large amount of genomic data, it is possible to obtain the frequency of each amino acid in the protein.
If a residue[2] shows a very large frequency compared to all other possible ones, we can therefore infer that the site has some kind of functional relevance.
Nevertheless, cooperativity plays an essential role between amino acids in the formation of secondary or tertiary structures, therefore it is more interesting to understand how different amino acids are linked between them. This can be done by observing if a mutation of one amino acid is correlated to the variation of another one.
For instance, when a residue is mutated and another one tends to show high mutation frequency, we can say that these two amino acids are co-evolving.
By operating in this way, it is possible to create a map of *sectors*: residues which can even be quite far, both in the sequence of the protein and even physically far from each other, which cannot mutate independently from each other,

[2]The residue is defined as the monomer of a generic chain: in all the cases here it is then an amino acid of the protein.

since their contemporary presence is necessary for the correct functionality of the protein itself.

Relation between sectors and allostery It has been shown experimentally (Reynolds, McLaughlin, and Ranganathan 2011) that there exists a strong statistical correlation between allosteric sites and sectors situated in the surface of DHFR.
The experimental analyses, in order to identify (latent) allosteric sites in DHFR consist of engineering a library of chimeric DHFR where, for each surface site, a light-sensitive module is inserted. Measuring then the variation of the activity of the enzyme due to the exposure to the light in a specific site will allow to assess whether the site can be considered to be allosteric or not.
Since allosteric sites and sectors show great superpositions, for future analyses we will make many comparisons to the sectors obtained by SCA in the previously cited paper.
One could ask why the direct comparison to experimental data is not performed. It has to be specified that both evolutionary and experimental results have their own drawbacks and limitations.
For instance, SCA data is exposed to statistical errors, and experimental data is only performed at the surface of the proteins. The analyses which will be present in later sections are related to properties of both solvent-exposed and non-solvent-exposed amino acids.
For this reason, we decided to make comparisons to sectors and not to direct experimental data, even if these analyses may be a bit more indirect.

1.4 Water and allostery

In the previously mentioned models for allostery, the hydration shell of the protein has been completely neglected. Different earlier works have found out that water outside the protein, or water confined inside of it, plays a role in the propagation of the allosteric signal (Buchli et al. 2013; Mackay and Wilson 1986). In the first paper, for example, the propagation of the allosteric signal seems to be completely controlled by the variation of water density around PMZ itself.
Some naïve mechanisms have been proposed (Mackay and Wilson 1986), according to which water could work as a proxy for the transmission of the

allosteric information.
For instance, the rearrangements of water dipoles, mediated by protein motions, could be able to immediately transmit the relevant information across sites which are quite far in the protein.
Nevertheless, it is not known yet if this behaviour is shown in all proteins, and how to predict with some kind of model how the mechanism takes place.
It is not yet clear if a characteristic behaviour of water around allosteric sites may be a cause or an effect of allostery itself.
It would be quite surprising if water is discovered to be a direct cause for allosteric regulation, due to the fact that the time-scales for water reorientation are much shorter (scale of picoseconds) than the ones related to conformational changes in the protein (millisecond scales).
There is not a clear answer for the link between water and allostery, and this work partially tries to investigate for the existence of a connection between hydration shell dynamics and allosteric regulation.
We will indeed verify if sector residues, mentioned earlier, display peculiar hydration dynamic properties.
If this were the case, generalizing this to all proteins, we would be able to identify allosteric sites just by looking at hydration dynamic properties of each protein.

1.5 Molecular dynamics

Molecular dynamics consists in integrating Hamilton's equations of motion for each particle contained in the system.
The tool which has been used in this work is GROMACS (Abraham et al. 2015), which includes all the methods described below. The integration is generally performed by means of the Velocity Verlet algorithm (Swope et al. 1982). The interactions cannot clearly be known exactly, as normally, at the atom-level, there are many quantum effects which have to be considered. In order to overcome this problem, these interactions are stored in the so-called *forcefields*, where quantum effects are taken into account by approximating them by simple interactions among atoms. It is not clearly possible to extract quantum properties by only using the forcefields, as everything will only involve classical mechanics, but these have proved to be quite effective in extracting observables which match to experiments.

Since we are just integrating Hamilton's equations, the energy is going to be conserved throughout the trajectory, but we have no control whatsoever on the temperature and the pressure.
These parameters are of critical importance in biological systems and it is necessary to have means of controlling them. For this reason thermostats and barostats have been developed.

Thermostats and barostats There are many possible algorithms which are employed to keep the temperature constant on average. These algorithms have different purposes and pros and cons depending on the physical quantities which have to be extracted from the trajectory.
The thermostat algorithm which is employed in our simulations is velocity rescaling(Bussi, Donadio, and Michele Parrinello 2007). The velocities of the particles are adjusted so that the average value of the temperature is the desired one. It has the advantage that the distribution of the velocities (and of the total kinetic energy) is the one which we would expect from the canonical ensemble.
For what concerns the barostat, we are using two different algorithms (for technical reasons): Berendsen's (Berendsen et al. 1984) and Parrinello-Rahman's (M. Parrinello and Rahman 1981). They consist in varying the volume of the cell inside which is system is contained, so that the average pressure can reach the desired value.

Restraints During the equilibration phases, the system is subject to large fluctuations. This happens because the solvent and the protein are initially at 0K and therefore the increase of the temperature, and therefore of the average velocity of the particles will have abrupt oscillations.
It is therefore necessary to add the so-called *restraints*, which consist in adding a strong force to all heavy atoms[3], so that the initial equilibrium structure is not destroyed completely.
The algorithm which is used to enforce these restraints is called LINCS (Hess et al. 1997).

Analysis The analyses of the trajectories have been performed by using MDAnalysis (Michaud-Agrawal et al. 2011; Gowers et al. 2016).

[3]By heavy atoms, we mean all atoms except hydrogen ones.

2 DHFR correlation analysis

2.1 Introduction

In this section we will analyze the correlations between the oscillations of the CC distance and the ones of the position of the Cα atoms of each residue. The distance between these two carbon atoms, shown in blue in Fig. (1), as said in the background section, is extremely important for the reaction. If a coupling between a residue and the distance between those two carbon atoms is present, we may expect that this site could display allosteric properties.
Indeed, a strong coupling between them would imply that a given site, if perturbated, would directly affect catalytic properties of the protein, which is exactly what we mean by allostery.
What we should then hope to find are residues which are far from the active site, and that also show some kind of coupling to the CC distance.
Once we identify them, we see if there is some kind of correspondence between them and residues identified as sectors, described in previous sections.
The open conformation of DHFR has been neglected, because in that conformation the CC distance is significantly bigger and there are water molecules which can get inside the active site, which considerably increase the reaction free energy barrier. We therefore do not observe significant results in that configuration.
We should make some clarifications regarding this kind of analysis: a small correlation (as defined earlier), between the residue and the CC distance will not imply that the given residue is not allosteric. This occurs because there are many different ways to probe experimentally an enzyme and, depending from the perturbation, different results could be found.
Moreover, we are only looking at equilibrium properties of a single conformation of DHFR, therefore it is not obvious that all relevant quantities to allostery will be present in our trajectories. The timescales we are sampling are indeed much shorter than the ones related to changes in conformation (which could also be induced through allosteric effects).
What we find is that the previously mentioned correlations are very small in value, but it is possible to observe that the residues which showed higher correlations to the CC distance are quite likely to be connected to sectors.
It therefore seems like that there is no direct coupling between the catalytic

activity of DHFR and the fluctuations of each Cα atom. Nevertheless, the ones which have a bigger coupling show to be directly linked to sectors.

2.2 Correlation

The correlation which has been chosen in this work, for simplicity, has been the highly employed Pearson correlation coefficient. It is computed in the following way:

$$c_{i,j} = \frac{< (\vec{r}_i - < \vec{r}_i >) \cdot (\vec{r}_j - < \vec{r}_j >) >}{\sqrt{< (\vec{r}_i - < \vec{r}_i >)^2) < (\vec{r}_j - < \vec{r}_j >)^2)}}, \quad (1)$$

where $\vec{r}_{i,j}$ indicate the positions of two atoms and the averages are performed over time. The systems we analyzed are assumed to be equilibrated, and the ergodic hypothesis may be assumed to hold.
The value of the correlation will then be, by definition, in-between +1 and -1.
The drawback of calculating correlations in this way is that two positions could be correlated, but their movements could happen on opposite planes. According to this definition, in these cases, the two positions would be uncorrelated. Moreover, this correlation assumes that our variables are linearly correlated and that the information is instantaneously propagated in time.
The reason why this correlation has been employed is because it is simple to implement and faster than other tools.
There are better ways to compute the correlations, and this work may probably underestimate the values of the correlations for many sites.
What could be employed, in order to solve the mentioned issues, are correlations defined through the estimations of mutual information, and therefore the probability distributions of the observables we are interested in (Kraskov, Stögbauer, and Grassberger 2004; Lange and Grubmüller 2005).

2.3 Analysis

The preparation of the system is described in appendix A.
We computed the correlations c_i between the deviation of an i-th Cα atom from its average position and the CC vector distance, in the following way,

just like defined in Eq. (1):

$$c_i = \frac{< (\vec{r}_i - < \vec{r}_i >) \cdot (\vec{d} - < \vec{d} >) >}{\sqrt{< (\vec{r}_i - < \vec{r}_i >)^2) < (\vec{d} - < \vec{d} >)^2)}}, \tag{2}$$

where $\vec{r}_i$ is the position of the Cα atom in each time frame, and $\vec{d}$ is the CC distance vector, defined as follows:

$$\vec{d} = \vec{r}_{C_{DHF}} - \vec{r}_{C_{NADPH}}, \tag{3}$$

where $\vec{r}_{C_{DHF/NADPH}}$ indicate the positions of the two carbon atoms involved in the reaction, described in previous sections.

We have chosen to adopt a vector distance in order to keep more information while averaging. We have indeed obtained that employing only the modulus of the vectors in the correlations would yield much smaller correlations.

From this correlation, we can identify the residues which show an absolute value of the correlation larger than a certain threshold. We tried different thresholds, and we chose the value of 0.06. A residue which has a c_i larger than this threshold will be later identified as 'highly correlated residue' (HCR). Clearly we mean *high* in a relative sense, as the obtained correlations are not high, in a strict sense.

As said before, our purpose is trying to understand if there are any residues which are more correlated than others and that are at large distances from the active site.

The constraint for the distance is important, because, as described before, allosteric regulation takes place between sites which are far from the active site.

To do so, we can then compute, for each Cα atom of each residue, its average distance from the position of the middle-point between the two carbon atoms involved in the reaction.

We then obtain the graph shown in Fig. (4), where we plotted the absolute value of the correlation, for each residue, as a function of the distance from the active site.

We can notice that most of the residues which show high correlations are indeed close to the active site, and we see that the correlation is decreasing as a function of the distance. This is indeed quite reasonable.

Nevertheless, we are able to identify many residues which have a quite large distance, bigger than 10 Å(this value is justified in the Correlation Matrix section) from the active site.
We can also visualize these residues by looking at the 3D structure of DHFR, to better see how they are distributed, as shown in Fig. (6).
We can indeed notice that that a lot of HCRs are close to the active site (shown at the center, where the ligands are represented), but there many which are quite far from it.
Given our correlations, our purpose is then to verify if the HCRs are close or show some kind of connectivity to the sector residues.

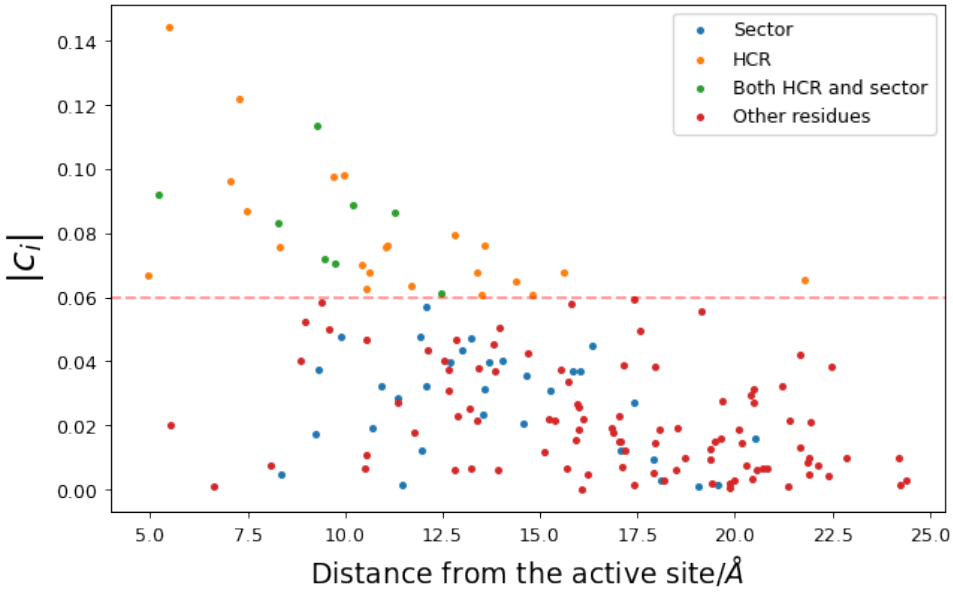


Figure 4: Correlations, computed as in Eq. (2), as a function of the distance from the active site. The latter is defined as the distance between the middle-point between the two relevant carbon atoms of the ligands, and the Cα atoms of each residue. These quantities are averaged over the trajectory.

Sector connectivity analysis In order to assess if an amino acid is connected to another one, we will compute the average distances among all Cα atoms of each residue. We will say that two residues are *connected* if the distance between their Cα atoms is less than 4Å.
The connectivity analysis will only be run over the HCRs and the sector residues: residues which do not belong to this category will not be involved in the connectivity.

Our assumption is that most of the HCRs are connected to sectors, in a network-like way. The reason why we look at network properties between HCRs is that correlation is more or less the same for close residues, therefore it is not really meaningful to look at connections between single amino acids: our networks will serve the purpose of a kind of coarse-graining.
If we also identify HCRs which are far from the active site, we may infer that these sites have some kind of allosteric significance.
These networks will be constructed iteratively in the following way: starting from each sector, we see if there exist HCRs which are connected to them, and in that case we add them to the network. Then, we proceed by checking if there are new HCRs which are connected to the ones added in the previous iteration, and we keep going until no HCR is added in the last step.
We can alternatively define the connectivity in the following way: we consider a graph where the vertices are only HCRs and sectors. there exists an edge between two vertices if the average distance between their associated Cα atom is less than 4 Å. We then say that an HCR is connected in a network-like way to a sector if there exists a path in the graph which leads its vertex to a sector vertex. We can notice that, among the 30 HCRs we identified (the identity of these residues is elicited in Appendix C), all of them except two appear to be connected to these networks, as shown in Fig. (5), obtaining a value around 93% for the connectivity. In order to verify if these thresholds, according to this model, would always produce a high connectivity, a p-value[4] study has been performed.
Given the correlations which we have obtained, we shuffled all the values and assigned them to random residues.
We then computed the connectivity for this random case and we repeated these calculations many times.
It has to be noted that, by doing these random permutations, we are adopting the hypothesis that each computed correlation can be equally distributed among each residue. This is probably not the case in proteins, as we should expect smaller correlations for residues which are far from the active site. We then obtain a histogram, in Fig. (7), through which we can notice that the event of obtaining 93% of HCRs connected to sectors is extremely unlikely.

[4]The p-value is a measure used to indicate the statistical significance of a hypothesis. The smaller it is, the less likely it is to measure the obtained value by chance, under a certain null hypothesis. In biology, a measure is considered to be significant when the p-value is smaller than 0.05.

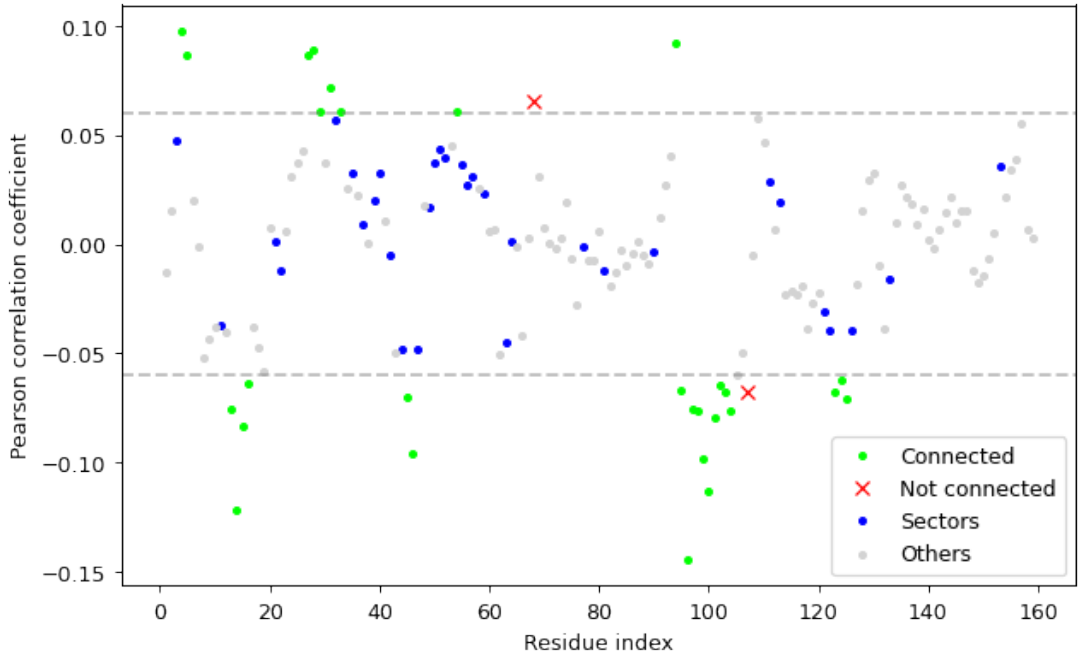


Figure 5: Pearson correlation, as in Eq. (2), as a function of the residue index. This means that we are numbering each amino acid in the protein, from 1 to 159: the protein is indeed a chain of amino acids. The chosen distance threshold for connectivity is 4Å.

Indeed, after 50,000 random permutations, it never occurs that the matching is higher or equal than 93%. We can then say that the p-value is less than 2×10^{-5}.

Assuming then that the available data is correct, this should reinforce the fact that the model is relevant and we can then infer that most HCRs, in DHFR, are close to these sectors.

We must clarify that, despite the fact that all HCRs are connected to sectors, the opposite is not true. There is not a one to one correspondence and this is also what occurs for residues with other properties (Reynolds, McLaughlin, and Ranganathan 2011).

Since many of the HCRs are close (distance less than 10 Å) to the active site, it may be interesting to perform this analysis only for the residues which are far from the active site.

By starting from the already performed connectivity analysis, we can consider only the residues which are further than a certain distance, for example 10 Å. This distance has been chosen because the scale over which the correlation as a function of the distance decreases by about two thirds is indeed about 5 Å,

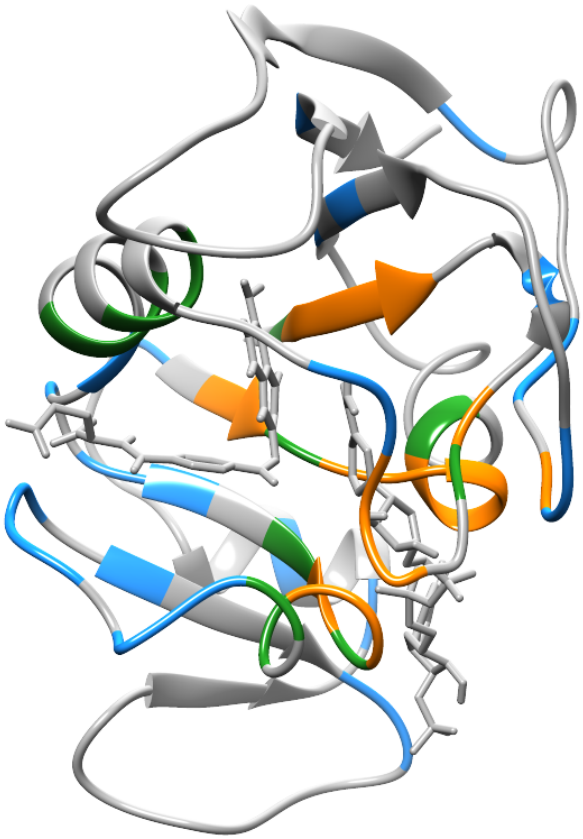


Figure 6: 3D visualization of DHFR in the closed configuration with highlighted HCRs (in orange), sectors (light blue) and residues which are both among the HCRs and sectors (green). It can be seen that the most HCRs are close to the active site, and almost always close or matching to sectors.

as shown later in Fig. (9).

By restricting our analysis to these residues, we obtain a matching of about 88%, and we again obtain an upper bound for the p-value equal to 2×10^{-5}. The obtained connectivity value is therefore significant and we conclude that, according to this analysis, most of the residues which show high correlations and are also far from the active site are connected, in a network-like way, to sectors.

From this analysis, we should not infer that each HCR, on its own, is more likely to be connected to a sector: the connections among the residues which show higher correlations are essential to be connected to sectors. The result is though hard to interpret from an experimental point of view. Starting from this analysis it could although be possible to identify allosteric pathways for DHFR, by analyzing the connections which are present among HCRs. By al-

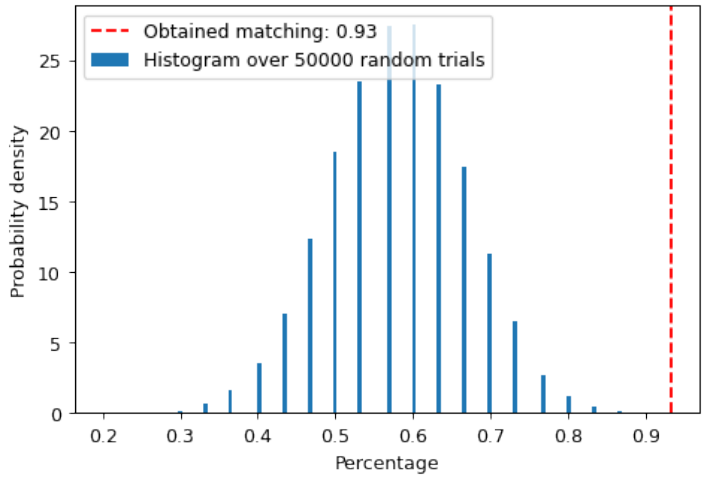
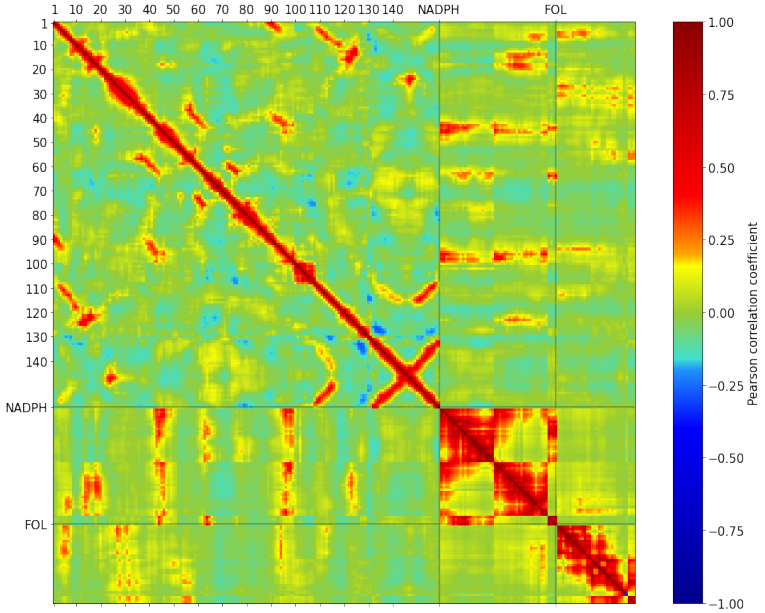


Figure 7: Histogram obtained after 50,000 random permutations of the correlations over the residues. We notice that the percentage value of 0.93 is never obtained over these attempts. We can then infer that the p-value is less than $1/50000 = 2 \times 10^{-5}$

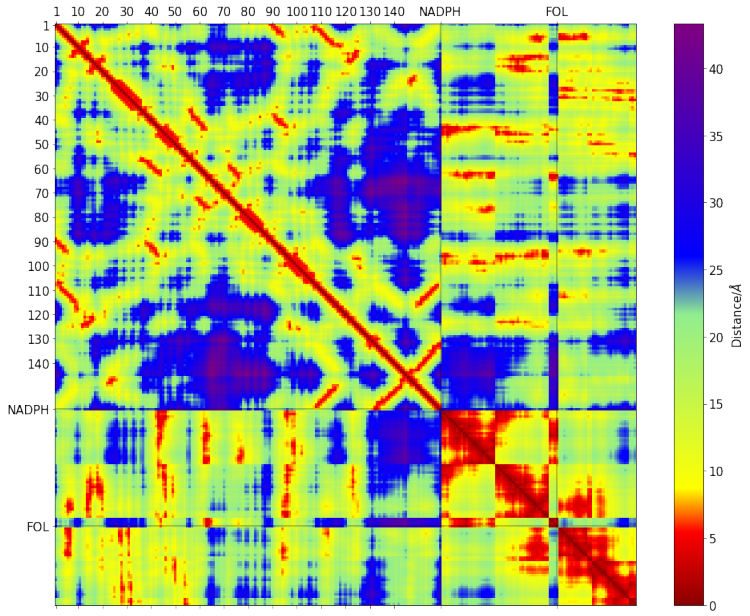
losteric pathway we would mean the path of connected HCRs starting from an HCR far from the active site towards the active site itself. Through this site, the information could propagate, and we partially could explain a mechanism for the propagation of the long-range interaction between an HCR and the catalytic activity of DHFR.
In the following section, we will address the question of the relevant spatial scale for the correlations. Indeed, we have stated here that a distance of 10 Å is said to be *large*, but we must motivate it.

2.4 Correlation matrix analysis

The purpose of this section is to study DHFR to identify a peculiar distance scale over which correlations decrease significantly, and to see if any explicit long-range correlations are present. Correlations will be computed for all Cα atoms of the residues, and the heavy atoms of the two ligands, in the same way defined in Eq. (1).
We will then compute all the correlations $c_{i,j}$, and we can then produce the correlation matrix, showed in Fig. (8a). We also compared this matrix to the one produced in literature, finding quite similar results (Boekelheide, Salomón-Ferrer, and Miller 2011).



(a) Correlation matrix



(b) Distance matrix

Figure 8: (a) Correlation matrix, obtained by considering the correlation of the deviation from the average position of all Cα atoms of the residues and all heavy atoms of the ligands, as in Eq. (1).
(b) Distance matrix, computed by considering the same atoms used in the correlation matrix, averaged over the trajectory.

We straightforwardly expect that atoms which are close to each other should display high correlations: in order to verify this we also plotted the distance

matrix, in Fig. (8b), where each entry $d_{i,j}$ is simply equal to the average distance between the i-th and j-th atom in question.
We can compare the two matrices and we notice that indeed there are very similar patterns between them.
In order to better analyze if there exist some residues which show a large correlation at large distances, we can perform a scatter plot for each couple of distances and the absolute value of the correlations $(d_{i,j}, |c_{i,j}|)$, which can be seen in Fig. (9). We obtain, as we might have expected, that correlations decay

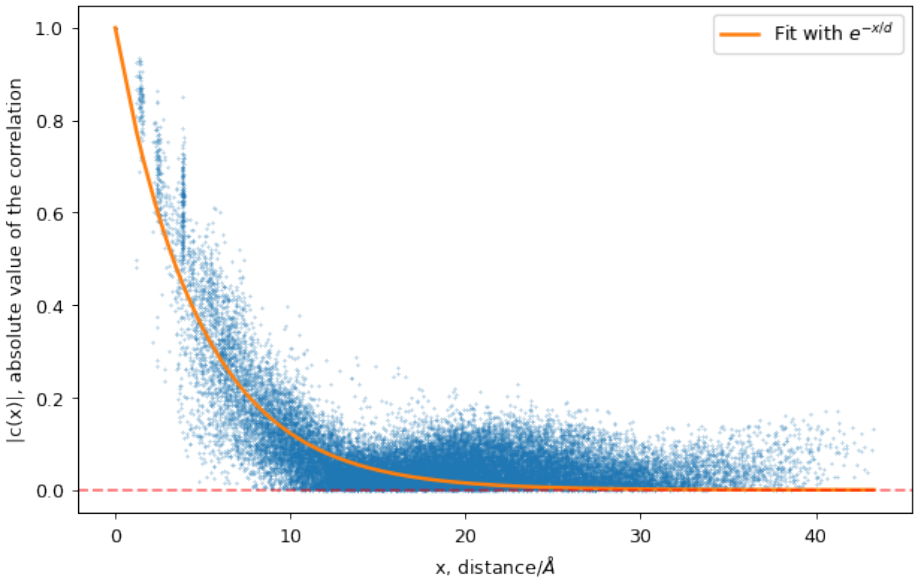


Figure 9: Scatter plot of the entries of the absolute value of the correlation and distance matrices, in Fig. (8). The trend appears to be exponentially decreasing and can then be fit with it (in orange). There are some large-distance correlations, but they are very small. The graphs make explicit how the correlations are most likely related only to close residues and we infer the typical length d over which these correlations decay.

as a function of the distance, and we can fit the points with an exponential in the following way:

$$c(x) = \mathrm{e}^{-\frac{x}{d}}, \tag{4}$$

where c is the correlation, x a given distance and d will be regarded as the typical length scale after which correlations start to become negligible (about one third), which has to be fit.
We have then obtained that d is around 4.8 Å.
This value justifies our previous distance threshold of 10 Å, which we used to

separate residues far from the active site and close to it.
We do not identify strong correlations at large distances, but some large distance weak correlations are displayed.

Principal component analysis To see if there are any relevant features which would allow us to understand if there exist any characteristics (specific of correlations) which differentiate HCRs from other residues, we decided to perform a Principal Component Analysis (PCA) of the correlation matrix. The first two principal components have about 40% of the explained variance, therefore a 2D plot should be able to catch many of the relevant features.
The projection of the correlation matrix over its two first principal components

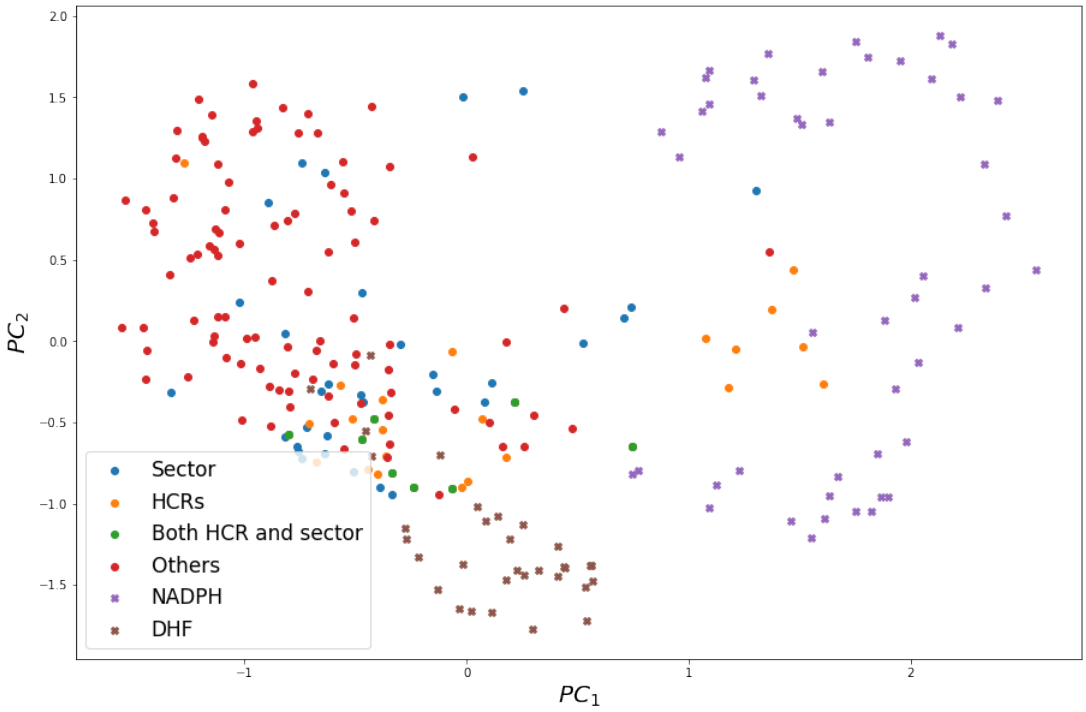


Figure 10: Principal component analysis over the first two principal components of the correlation matrix. The criteria for the selection of HCRs are the same ones which were mentioned in the Connectivity Analysis.

is shown in Fig. (10).
We can notice that the ligands and the residues of the protein seem to occupy quite separate regions. We see indeed that there is a cluster of mostly uncorrelated residues, a cluster for each of the ligands and an intermediate behaviour which is associated to HCRs.

From the figure, it may seem like that HCRs display a similar correlation behaviour to the one of the ligands, and especially the cofactor (in purple).
There are indeed some residues, which are mostly HCRs, which appear to behave quite similarly to it, since they are very close to it in the graph.
Moreover, in principle, this should not necessarily be related to the fact that those HCRs are close to NADPH. Indeed all the distances of those residues are higher than 5 Å, and only two are larger than 10 Å, the previously chosen threshold to identify HCRs far from the active site.
From this analysis, we may infer that HCRs have in general (global) correlation properties which are more similar to the ones of the ligands, and different from other residues.This should not necessarily be obvious, because the way in which we defined the HCRs is quite less informative than computing the full correlation matrix, where we have computed correlations for each pair of atom. It has to be said though that most HCRs are close to non-HCRs.
For this reason, it does not seem to be possible to extract an unequivocal reason which explains this behaviour, at least from this analysis.
Finally, we notice that sectors, even less than HCRs, are not well separated from the rest of the residues. It seems like that the sectors do not display any peculiar correlation behaviour.

3 DHFR hydration shell analysis

3.1 Introduction

After having performed a correlation study, we will now focus on some properties related to water hydration shell dynamics (some details of the MD simulations are presented in Appendix B.). As mentioned earlier, it is not clear if water plays a role or not in allosteric regulation.
In this section, we will determine if the residues involved in sectors display characteristic hydration dynamic properties.
What we will have to do is then to extract the water reorientational times in the shell around the protein. These observables have been obtained by using a tool developed at ENS by Élise Duboué-Dijon, using previously developed strategies (Fogarty and Laage 2014).
We have extracted and analysed these quantities for both the open and closed conformations of DHFR. In the following sections we have then decided to present the results for the open configuration of DHFR, but, for the reasons explained earlier, we do not expect real physical relevance, and we just included it as a reference for comparison.

3.2 Water hydration dynamics

Water is essential for biological systems and its dynamics is capable of influencing the structure and function of some biomolecules (Ball 2007; Levy and Onuchic 2006).
Water molecules are forming hydrogen bonds between one another, but water is not a static solvent. This implies that it continuously breaks and forms new hydrogen bonds, within a picosecond timescale.
The presence of a protein (or a generic biomolecule) in water will though perturb water dynamics, in the sense that water reorientation is much slower around it.
One well known model has been developed in the past (Frank and Evans 1945): the iceberg model. According to it, the presence of hydrophobic sites would reduce water configurations, because it limits its ability to form hydrogen bonds. This would in turn translate into a great loss of entropy of water because fewer configurations are for them available. According to this model, some icebergs are formed around solutes in water, where water behaves as if it were frozen.

It has been shown (Laage, Stirnemann, and Hynes 2009) that this is not really occurring: water molecules can move around the protein and we can measure both experimentally and also by using simulations their reorientational times. Water dynamics is therefore slowed around the protein, but it is not frozen.
There are two main static properties which influence water dynamics around a protein: the geometry of the protein and its chemical properties.
Indeed, the protein is quite inhomogeneous, both geometrically and chemically, and for this reason water dynamics shows very different characteristics around it, especially if there are pockets[5] in the protein where water is allowed to enter.
Moreover, since the protein does not fluctuate in the same way around its surface, there will also be dynamical properties which in turn influence water dynamics.
In this work, we will limit ourselves to the analysis of the first hydration shell of the protein.
The definition of the first hydration shell is an arbitrary concept. In general, the hydration shell is constituted of the water molecules which are more perturbed by the presence of a biomolecule. To define it, since we are using Molecular Dynamics tools to analyze our trajectories, we can introduce some geometric constraints related to the distance between water molecules and atoms of the protein.
In order to keep chemical differences into account, we will classify each atom into three classes: hydrogen bond donor, hydrogen bond acceptor or hydrophobic[6]: each of these classes will have a different distance threshold, usually obtained by considering radial distribution functions, as described in literature (Fogarty and Laage 2014).
Once we identify the water molecules of interest, we will assign them to the site to which they are close, and there are some technicalities which are also discussed in the aforementioned paper.
It is then possible to compute the time autocorrelation function related to the orientation of the OH-bond vectors[7] of each water molecule, in the following

[5] A protein pocket is a generic cavity inside the surface of a protein.

[6] A hydrogen bond donor is an atom which is attached to a strong electronegative atom: it will then have a positive partial charge around it so that it can interact with other more electronegative atoms. For example the H atoms in water tend to form hydrogen bonds with oxygen, which is a hydrogen bond acceptor.

[7] Each water molecule will have two OH vectors: they are the two vectors starting from the oxygen atom and pointing towards one of the other hydrogen atoms.

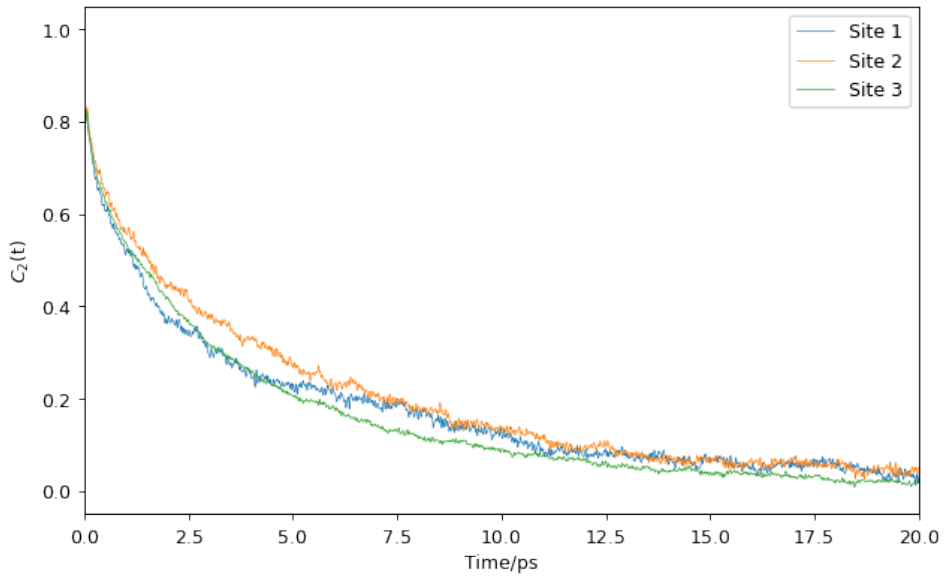
way:

$$C_2(t) = \langle P_2(\vec{u}(0) \cdot \vec{u}(t)) \rangle \,, \tag{5}$$

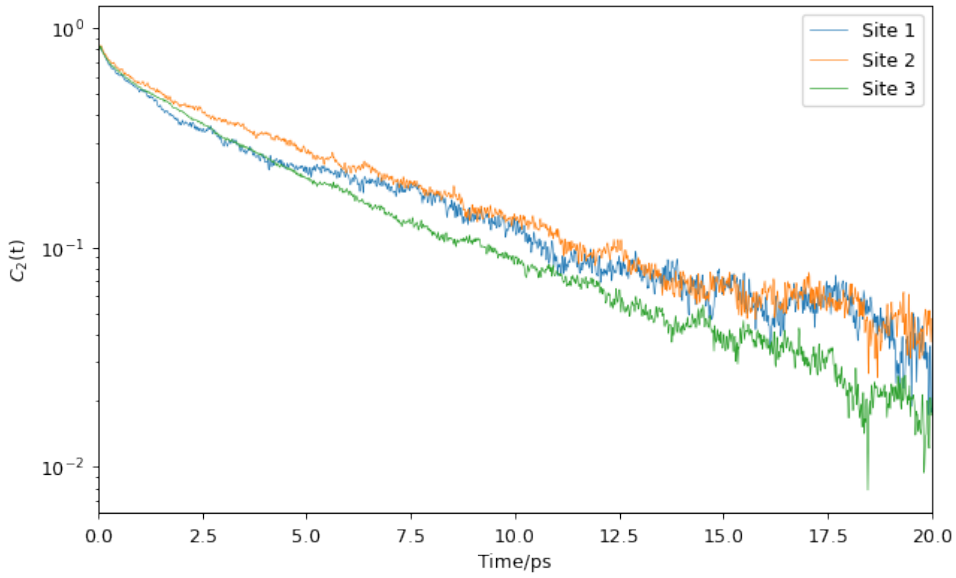
where P_2 indicates the second order Legendre polynomial, and $\vec{u}$ is the OH-bond vector for each water molecule.
The second order Legendre polynomial is adopted because it can be related to experimentally accessible quantities.
These correlation functions decay over time, as seen in Fig. (11a), and the main goal will be to estimate the time over which the relaxation occurs.
The decay is non-monoexponential, as can better be seen in the log-y scale in Fig. (11b).This behaviour is related to initial fast (sub-picosecond) librational relaxation of the OH vector. For this reason, in order to obtain the reorientational time, we will fit these curves within an interval between 2 and 10 picoseconds. This is done to avoid fitting the intial multi-exponential behaviour and to avoid the strong noise which is present in the final parts of the curve. By estimating the relaxation time for each OH-bond vector around the protein, we can obtain a map of reorientational times for each site in the protein.

3.3 Analyses

We can notice that the two conformations of DHFR have qualitatively similar reorientational times, but many sites seem to behave quite differently, as it can be seen from Fig. (12).
This can probably explained by the fact that, despite the fact that both structures are composed of the same amino acids, they have quite different fluctuations for some key observables, like the Met20 loop. These fluctuations will considerably affect reorientation of water around these sites.
Moreover, the two conformations will in some cases expose different amino acids to water, and this will also affect water dynamics. What can be done for the OH bond reorientational times is to see how they are distributed. We then plotted the histograms in order to infer the distribution, for both conformations Fig. (13). We reproduce similar distributions to the ones obtained in literature (Sterpone, Stirnemann, and Laage 2012). The distributions of the reorientational times of the two conformations appear to be quite similar, and a peak in the range of 2-3ps is displayed.



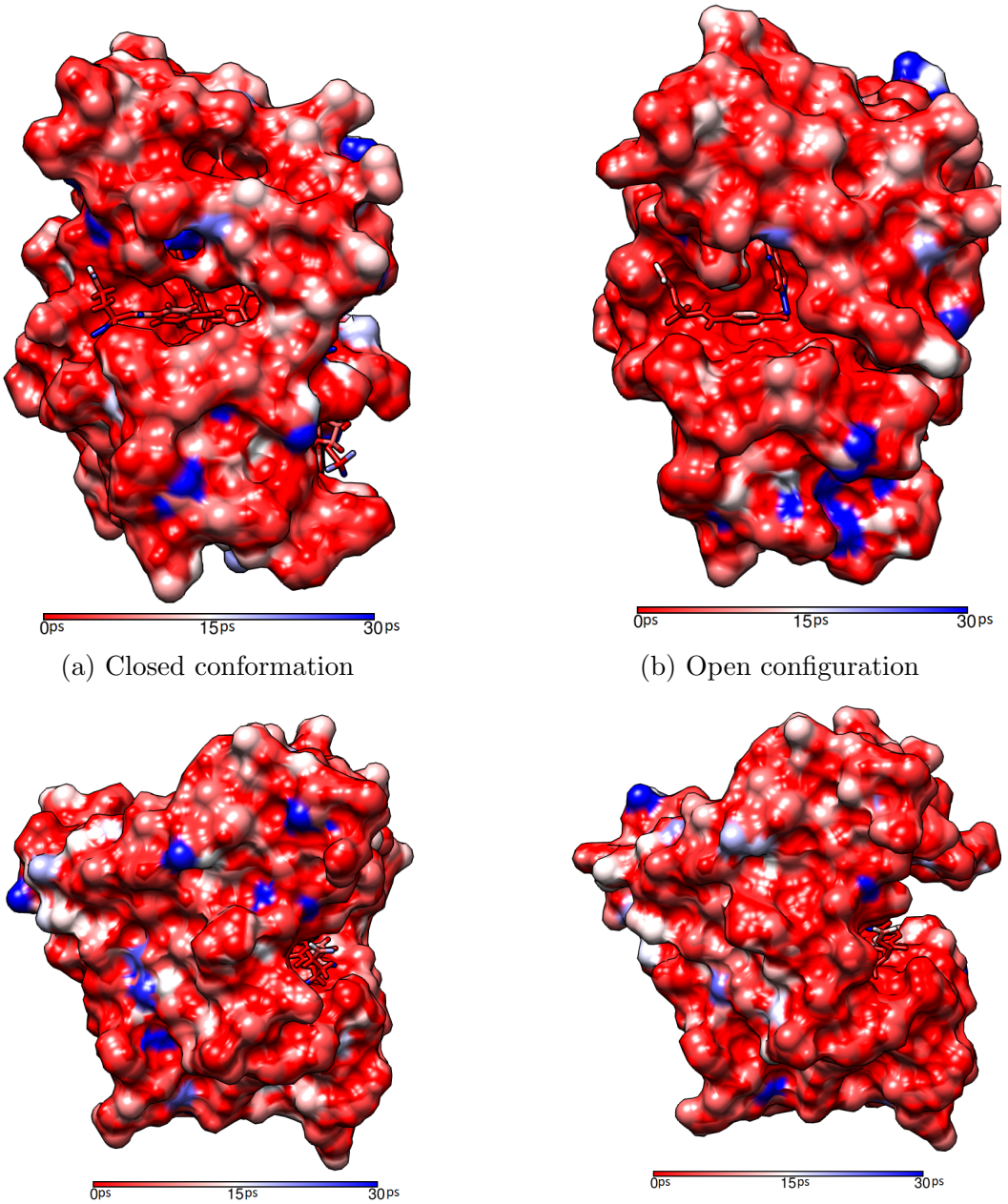
(a)



(b)

Figure 11: Autocorrelation function $C_2(t)$, as in Eq. (5), for three arbitrary sites in the closed configuration of DHFR, (a) in linear scale, (b) in log-y scale. By fitting intermediate intervals of these functions, for example in-between 2-10ps, it is possible to extract OH vector reorientational times.

It is indeed known (Laage, Elsaesser, and Hynes 2017) that this peak is generally caused by the local topography of the interface between the protein and water, which limits the rearrangements of the H-bonds.
The tail of the distribution instead generally arises from water molecules which



(a) Closed conformation

(b) Open configuration

(c) Closed configuration, lateral view

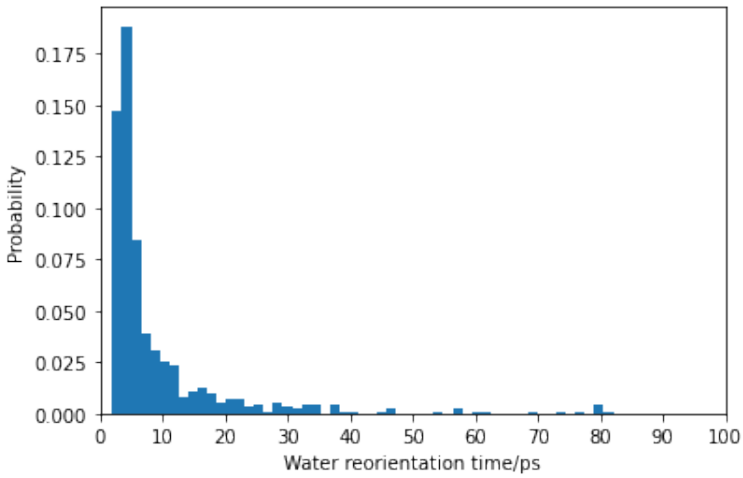
(d) Open configuration, lateral view

Figure 12: Water colour map of the reorientational times (in picoseconds) for the closed (a),(c) and open (b),(d) configurations of DHFR, seen from two perspectives.

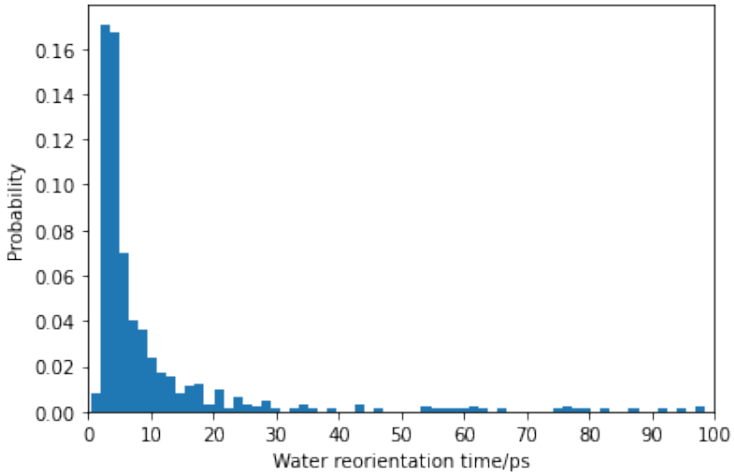
are buried inside pockets. This also seems to be the case, since the tails appear to be longest in the open conformation, where water can also get in the active site.

If we perform a Smirnov-Kolmogorov test[8], we obtain a small value of 0.065, but with a p-value of 0.15. It is therefore not clear if the two distributions are the same or not, despite appearing qualitatively quite similar.
The peak in the distribution appears to be slightly different for the two conformations: the closed conformation displays a maximum in the distribution at about $3.5ps$, while $1.9ps$ is the one found for the open one.
Since we want to verify if sector residues display special hydration dynamic properties, we could think that some characteristics would be evident even from the differences in the distributions of reorientational times of sector and non-sector sites. The distributions are showed in Fig. (14).
We notice that for both distributions, and especially for the closed one, the peaks seem to be a tiny bit higher for larger reorientational times.
In order to verify if this is quantitatively reproduced, we performed two different Smirnov-Kolmogorov tests.
From this, we do not notice a large quantitative difference: we obtain, in both cases, a value around 0.1, but p-values are quite high, so the measure is not significant and we cannot infer much from it.
We therefore cannot conclude from the distributions that sites belonging to sector residues display characteristic hydration dynamic properties.
Another possible question we might ask is if the slowest sites are physically located close to sectors. In order to verify if such correspondences exist, we will therefore perform a similar connectivity analysis to the one performed in the previous section for HCRs. Before running the connectivity analysis, we average the obtained reorientational times for each single amino acid. Therefore each residue will be associated to a single reorientational time for water around it.
Very naively, we then sum the reorientational times of each site in the single residue and we then divide by the number of solvent-exposed sites present in it.
The same connectivity analyses which were performed for the HCRs are now performed for the aforementioned residues.
In this case we identify a subset of slow sites, the residues around which water reorientational time is highest (indicated in Appendix C). We set a threshold

[8]The Smirnov-Kolmogorov test is a statistical test which allows to compare two samples and it will return a value related to how likely it is that these two samples have been generated by the same probability density function.



(a) Closed configuration



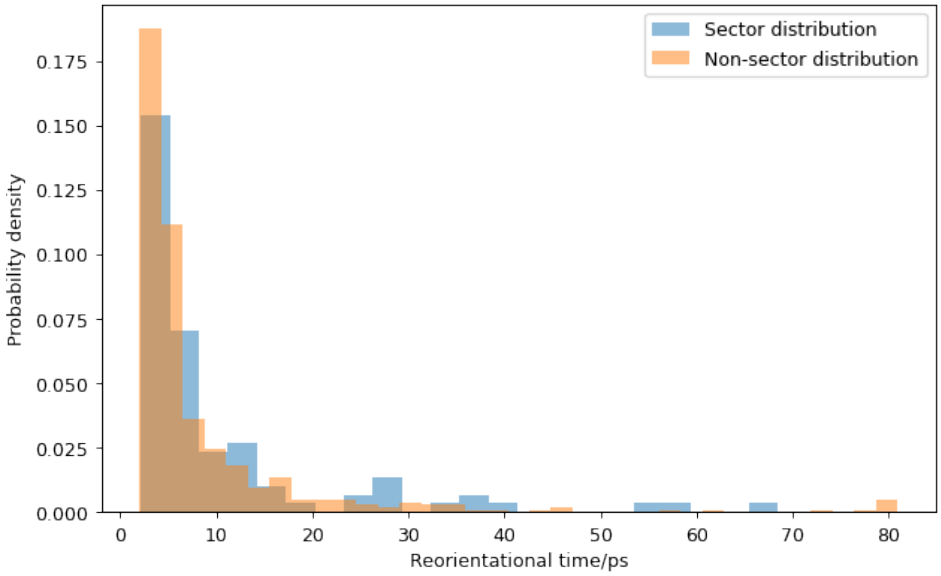
(b) Open configuration

Figure 13: Distribution of the reorientational times for water OH vectors in the closed (a) and open (b) conformations of DHFR, for the first hydration shell. There are not qualitative differences between the two distributions and they both follow the behaviour expected from literature.

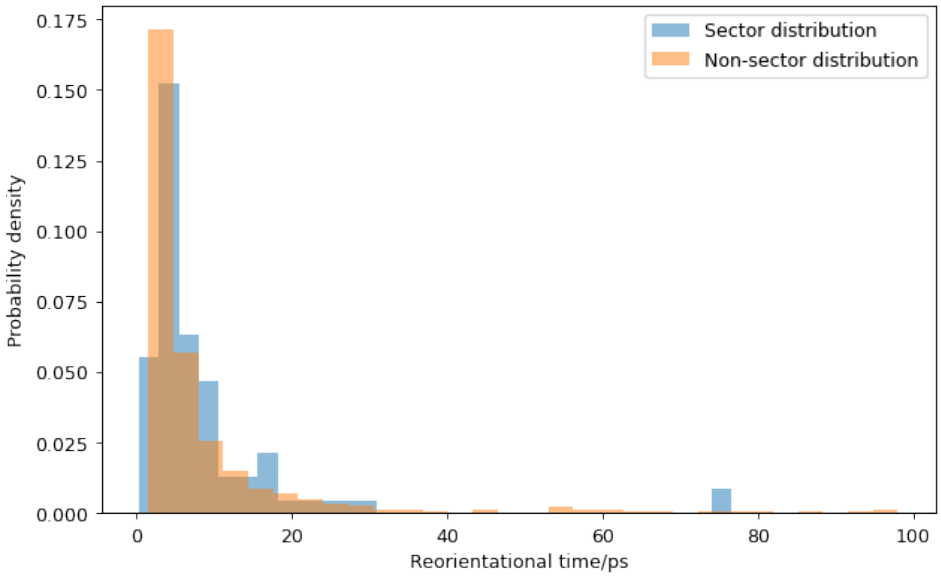
of 35ps in order to classify them as such. By using the same algorithm, with the same distance threshold of 4 Å, we now obtain that about 76% and 55% of the sites are indeed connected to sectors, for the closed and the open configuration respectively.

In reality, we can notice that network connectivity is actually not important in this case: all slow sites are directly connected to sector ones, without the need of other intermediate connections among them.

This is reasonable, since slow sites are only in the surface of the protein, and



(a) Closed configuration



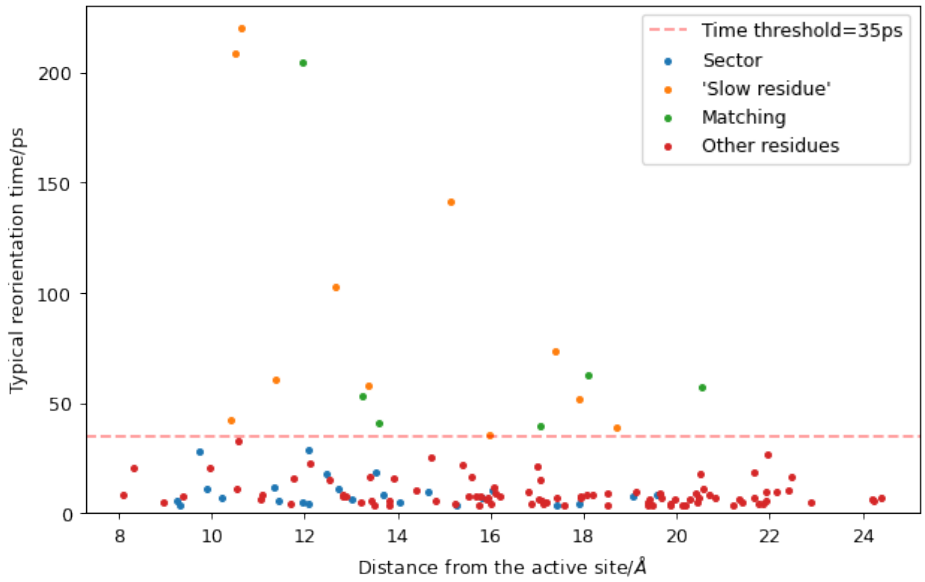
(b) Open configuration

Figure 14: Comparison between the distribution of reorientational times of water OH vectors for sites belonging to sectors (blue) and non-sector sites (orange), in the closed (a) and open (b) conformations of DHFR. Apparently, it seems that the closed conformation shows some differences in the behaviour between sector-sites and non-sector ones, especially in the tails of the distributions.

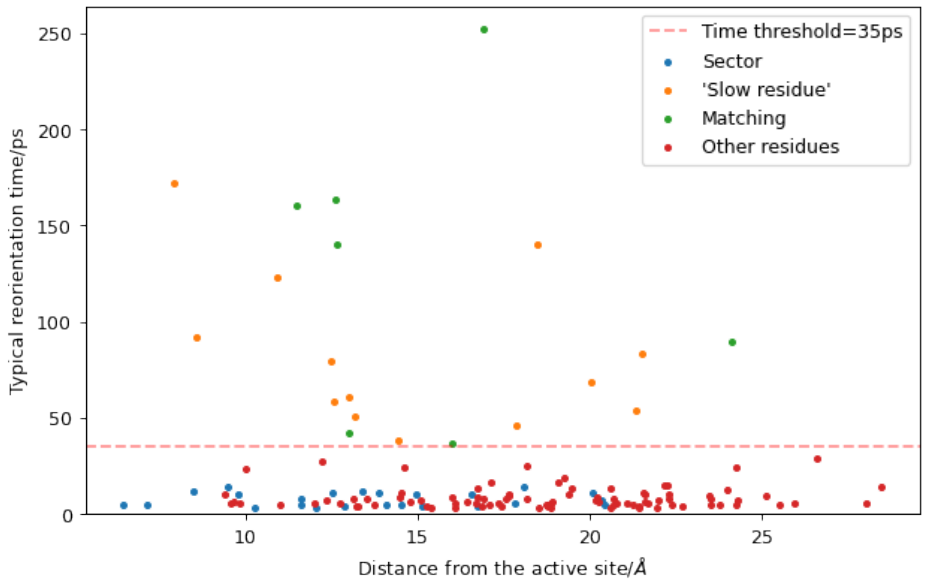
it is more unlikely for them to be close to one another.

In the correlation study we made, it was straightforward that correlated residues

should be close to other ones, since we expect that close residues have similar correlations. On the other hand, in this case, the chemical properties of each amino acid are relevant. Since the surface of the protein is chemically quite inhomogeneous, we do not expect that slow sites are close to one another.
Results and p-values for this section are therefore obtained by assuming that no network of slow sites is formed: for each slow site, we only look if it is directly close to a sector.
Again, we verify that our results are unlikely to be obtained by chance, measuring the statistical significance through the p-value. The latter has been estimated by permutating the 'coarse-grained' times for the solvent-exposed residues, and by repeating the same calculations many times.
We obtain a p-value of 0.03 for the closed configuration, making the measure significant.
Moreover, we can also notice that these sectors are now not close to the active site: for both the closed and open conformation, the strong majority (all of them, in the closed configuration) of all slow sites shows a distance larger than 10 Å from the active site, as can be seen in Fig. (15) and partially from Fig. (16). We can then restrict our analysis to sites which show a distance larger than 10 Å from the active site. We obtain the same percentage of slow sites which are connected to sectors, for the closed configuration, and the p-value is estimated to be less than 5×10^{-5}. This connectivity is therefore even more significant than the one obtained for all slow sites.
We have therefore obtained that most slow sites are directly connected to sectors: it would be interesting to verify this hypothesis both experimentally and with other kinds of proteins, to verify if this happens in general or if it is something peculiar of DHFR.
This is important because since a large amount of slow sites is connected to sectors, it would be possible to have some hints on where allosteric sites are situated just by performing molecular dynamics simulations.



(a) Closed configuration



(b) Open configuration

Figure 15: Average reorientational times around each surface residue as a function of the distance from the active site, computed as described in the previous section. We notice that there is not a clear correlation between the reorientational times and the distance from the active site, differently from what was observed for the correlations.

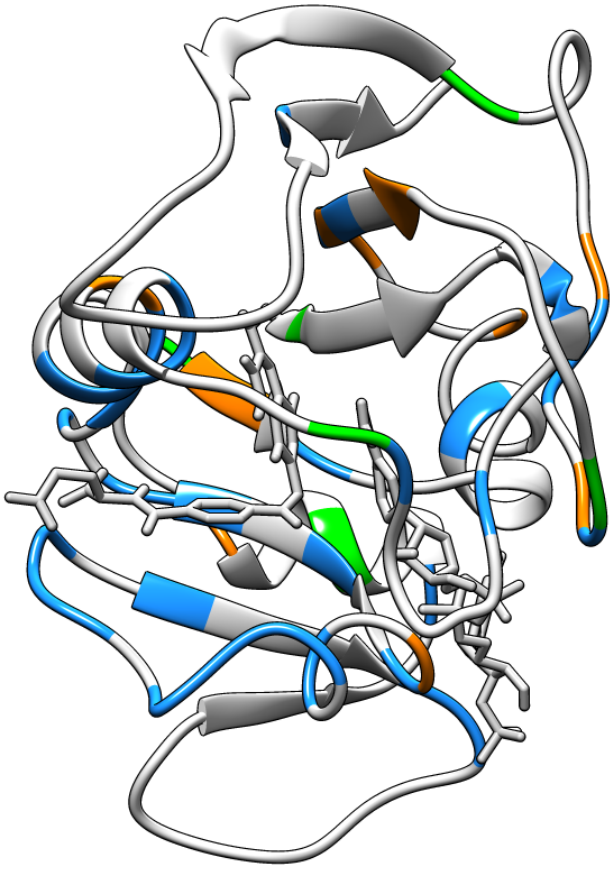


Figure 16: 3D representation of DHFR in the closed conformation. The sites around which the water reorientational times are highest (orange) are compared to sectors (light blue), in the case when they match, they are represented in green. The open configuration has not been represented because it shows low connectivity with large p-values: the measure is not significant.

4 Conclusions

In this work we have analysed some equilibrium properties of DHFR, such as correlation properties between the positions of amino acids and the CC distance which governs its catalytic properties.
We wanted to verify if there existed a matching between sector residues, obtained from evolutionary data, and residues which showed higher correlations. Even if the obtained correlations are not quite large, we showed that the strong majority (86%) of residues which displayed higher correlations (HCRs), and which are far from the active site, are connected, in a network-like way, to sector residues. This implies that there may exist a relation between networks of HCRs and catalytic activity, but it has to be clarified that there seems not to be a direct link between single HCRs and sectors: the matching without network-like connections would be quite small. Additional analyses of the HCRs and their connections could be useful to identify allosteric pathways.
We have also analysed the first hydration shell of DHFR, by looking at the reorientational times of OH vectors of water. Our purpose was to verify if sector residues had peculiar hydration dynamics, but it has not been possible to infer it from our data. We therefore tried to see if sector residues are directly connected to residues around which OH vectors have largest reorientational times (called slow sites). We noticed that slow sites far from the active site are, most often (76%), directly connected to sectors. This result has stronger implications than the one obtained for HCRs, because, unlike the latter, it provides a direct link between slow sites and sectors, even if the matching percentage which was obtained is smaller. If this is verified, we could have hints on where to start looking experimentally for allosteric sites in a protein.
Moreover, it would be interesting to understand the molecular mechanisms which would lead water to have larger reorientational times around sectors.
It also has to be said that the two analyses which we have performed, identify different residues, except for two common ones. Additional analyses show that the two subsets of residues do not even appear to be connected.
In this work we have only looked at equilibrium properties of DHFR and its hydration shell, and we never investigated on **how** the information travels across the protein.
A very promising way to continue the project could be to use other techniques, such as Nonequilibrium Molecular Dynamics simulations (Oliveira et al. 2021),

which would allow us to directly verify if, applied a given perturbation, the information travels across the protein or through the reorientation of water around the surface of the protein.
These studies could provide a clear answer to the question of a possible role of water in allosteric transitions.

References

Abraham, Mark James et al. (Sept. 2015). "GROMACS: High performance molecular simulations through multi-level parallelism from laptops to supercomputers". en. In: *SoftwareX* 1–2, pp. 19–25. DOI: 10.1016/j.softx.2015.06.001. URL: http://dx.doi.org/10.1016/j.softx.2015.06.001.

Ahmed, Mostafa H., Mohini S. Ghatge, and Martin K. Safo (2020). *Hemoglobin: Structure, Function and Allostery*. DOI: 10.1007/978-3-030-41769-7_14. URL: http://dx.doi.org/10.1007/978-3-030-41769-7_14.

Ball, Philip (Dec. 2007). *Water as an Active Constituent in Cell Biology*. en. DOI: 10.1021/cr068037a. URL: http://dx.doi.org/10.1021/cr068037a.

Berendsen, H. J. C. et al. (Oct. 1984). "Molecular dynamics with coupling to an external bath". en. In: *The Journal of Chemical Physics* 81.8, pp. 3684–3690. DOI: 10.1063/1.448118. URL: http://dx.doi.org/10.1063/1.448118.

Boekelheide, Nicholas, Romelia Salomón-Ferrer, and III Miller Thomas F. (Sept. 2011). *Dynamics and dissipation in enzyme catalysis*. en. DOI: 10.1073/pnas.1106397108. URL: http://dx.doi.org/10.1073/pnas.1106397108.

Bowerman, S. and J. Wereszczynski (2016). *Detecting Allosteric Networks Using Molecular Dynamics Simulation*. DOI: 10.1016/bs.mie.2016.05.027. URL: http://dx.doi.org/10.1016/bs.mie.2016.05.027.

Brooks, B.R., C.L. Brooks III, and A.D. Mackerell Jr. et al. (2009). "CHARMM: The Biomolecular Simulation Program". In: *J. Comput. Chem.* 30. PMC2810661, 1545–1614.

Buchli, Brigitte et al. (July 2013). "Kinetic response of a photoperturbed allosteric protein". en. In: *Proceedings of the National Academy of Sciences* 110.29, pp. 11725–11730. DOI: 10.1073/pnas.1306323110. URL: http://dx.doi.org/10.1073/pnas.1306323110.

Bussi, Giovanni, Davide Donadio, and Michele Parrinello (Jan. 2007). "Canonical sampling through velocity rescaling". en. In: *The Journal of Chemical Physics* 126.1, p. 014101. DOI: 10.1063/1.2408420. URL: http://dx.doi.org/10.1063/1.2408420.

Changeux, Jean-Pierre (2012). "Allostery and the Monod-Wyman-Changeux Model After 50 Years". In: *Annual Review of Biophysics* 41.1. PMID: 22224598, pp. 103–133. DOI: 10.1146/annurev-biophys-050511-102222. eprint: https://doi.org/10.1146/annurev-biophys-050511-102222. URL: https://doi.org/10.1146/annurev-biophys-050511-102222.

Feher, Victoria A et al. (Apr. 2014). "Computational approaches to mapping allosteric pathways". en. In: *Current Opinion in Structural Biology* 25, pp. 98–103. DOI: 10.1016/j.sbi.2014.02.004. URL: http://dx.doi.org/10.1016/j.sbi.2014.02.004.

Fogarty, Aoife C. and Damien Laage (Feb. 2014). "Water Dynamics in Protein Hydration Shells: The Molecular Origins of the Dynamical Perturbation". en. In: *The Journal of Physical Chemistry B* 118.28, pp. 7715–7729. DOI: 10.1021/jp409805p. URL: http://dx.doi.org/10.1021/jp409805p.

Frank, Henry S. and Marjorie W. Evans (Nov. 1945). *Free Volume and Entropy in Condensed Systems III. Entropy in Binary Liquid Mixtures; Partial Molal Entropy in Dilute Solutions; Structure and Thermodynamics in Aqueous Electrolytes*. en. DOI: 10.1063/1.1723985. URL: http://dx.doi.org/10.1063/1.1723985.

Garcia, Hernan G. et al. (2011). *Thermodynamics of Biological Processes*. DOI: 10.1016/b978-0-12-381268-1.00014-8. URL: http://dx.doi.org/10.1016/B978-0-12-381268-1.00014-8.

Gomez, Axel et al. (May 2022). *Water Diffusion Proceeds via a Hydrogen-Bond Jump Exchange Mechanism*. en. DOI: 10.1021/acs.jpclett.2c00825. URL: http://dx.doi.org/10.1021/acs.jpclett.2c00825.

Gowers, Richard J. et al. (2016). "MDAnalysis: A Python Package for the Rapid Analysis of Molecular Dynamics Simulations". In: *Proceedings of the 15th Python in Science Conference*. Ed. by Sebastian Benthall and Scott Rostrup, pp. 98–105. DOI: 10.25080/Majora-629e541a-00e.

Grover, Ashok Kumar (2013). *Use of Allosteric Targets in the Discovery of Safer Drugs*. en. DOI: 10.1159/000350417. URL: http://dx.doi.org/10.1159/000350417.

Guo, Jingjing and Huan-Xiang Zhou (Feb. 2016). *Protein Allostery and Conformational Dynamics*. en. DOI: 10.1021/acs.chemrev.5b00590. URL: http://dx.doi.org/10.1021/acs.chemrev.5b00590.

Halabi, Najeeb et al. (Aug. 2009). "Protein Sectors: Evolutionary Units of Three-Dimensional Structure". en. In: *Cell* 138.4, pp. 774–786. DOI: 10.

1016/j.cell.2009.07.038. URL: http://dx.doi.org/10.1016/j.cell.2009.07.038.

Hess, Berk et al. (Sept. 1997). "LINCS: A linear constraint solver for molecular simulations". en. In: *Journal of Computational Chemistry* 18.12, pp. 1463–1472. DOI: 10.1002/(sici)1096-987x(199709)18:12<1463::aid-jcc4>3.0.co;2-h. URL: http://dx.doi.org/10.1002/(SICI)1096-987X(199709)18:12%3C1463::AID-JCC4%3E3.0.CO;2-H.

Kraskov, Alexander, Harald Stögbauer, and Peter Grassberger (June 2004). "Estimating mutual information". en. In: *Physical Review E* 69.6. DOI: 10.1103/physreve.69.066138. URL: http://dx.doi.org/10.1103/PhysRevE.69.066138.

Laage, Damien, Thomas Elsaesser, and James T. Hynes (Mar. 2017). *Water Dynamics in the Hydration Shells of Biomolecules*. en. DOI: 10.1021/acs.chemrev.6b00765. URL: http://dx.doi.org/10.1021/acs.chemrev.6b00765.

Laage, Damien, Guillaume Stirnemann, and James T. Hynes (Feb. 2009). *Why Water Reorientation Slows without Iceberg Formation around Hydrophobic Solutes*. en. DOI: 10.1021/jp809521t. URL: http://dx.doi.org/10.1021/jp809521t.

Lange, Oliver F. and Helmut Grubmüller (Dec. 2005). "Generalized correlation for biomolecular dynamics". en. In: *Proteins: Structure, Function, and Bioinformatics* 62.4, pp. 1053–1061. DOI: 10.1002/prot.20784. URL: http://dx.doi.org/10.1002/prot.20784.

Leitner, David M., Changbong Hyeon, and Korey M. Reid (June 2020). *Water-mediated biomolecular dynamics and allostery*. en. DOI: 10.1063/5.0011392. URL: http://dx.doi.org/10.1063/5.0011392.

Levy, Yaakov and José N. Onuchic (June 2006). *WATER MEDIATION IN PROTEIN FOLDING AND MOLECULAR RECOGNITION*. en. DOI: 10.1146/annurev.biophys.35.040405.102134. URL: http://dx.doi.org/10.1146/annurev.biophys.35.040405.102134.

Liu, C. Tony et al. (June 2013). *Functional significance of evolving protein sequence in dihydrofolate reductase from bacteria to humans*. en. DOI: 10.1073/pnas.1307130110. URL: http://dx.doi.org/10.1073/pnas.1307130110.

Lockless, Steve W. and Rama Ranganathan (Oct. 1999). "Evolutionarily Conserved Pathways of Energetic Connectivity in Protein Families". en. In:

Science 286.5438, pp. 295–299. DOI: 10.1126/science.286.5438.295. URL: http://dx.doi.org/10.1126/science.286.5438.295.

Mackay, Donald H. J. and Kent R. Wilson (Dec. 1986). *Possible Allosteric Significance of Water Structures in Proteins*. en. DOI: 10.1080/07391102.1986.10506364. URL: http://dx.doi.org/10.1080/07391102.1986.10506364.

Maffucci, Irene, Damien Laage, Fabio Sterpone, et al. (July 2020). *Thermal Adaptation of Enzymes: Impacts of Conformational Shifts on Catalytic Activation Energy and Optimum Temperature*. en. DOI: 10.1002/chem.202001973. URL: http://dx.doi.org/10.1002/chem.202001973.

Maffucci, Irene, Damien Laage, Guillaume Stirnemann, et al. (2020). "Differences in thermal structural changes and melting between mesophilic and thermophilic dihydrofolate reductase enzymes". en. In: *Physical Chemistry Chemical Physics* 22.33, pp. 18361–18373. DOI: 10.1039/d0cp02738c. URL: http://dx.doi.org/10.1039/d0cp02738c.

McCormick, James W et al. (June 2021). *Structurally distributed surface sites tune allosteric regulation*. en. DOI: 10.7554/elife.68346. URL: http://dx.doi.org/10.7554/eLife.68346.

Michaud-Agrawal, Naveen et al. (Apr. 2011). "MDAnalysis: A toolkit for the analysis of molecular dynamics simulations". en. In: *Journal of Computational Chemistry* 32.10, pp. 2319–2327. DOI: 10.1002/jcc.21787. URL: http://dx.doi.org/10.1002/jcc.21787.

Motlagh, Hesam N. et al. (Apr. 2014). *The ensemble nature of allostery*. en. DOI: 10.1038/nature13001. URL: http://dx.doi.org/10.1038/nature13001.

Nagae, Takayuki, Hiroyuki Yamada, and Nobuhisa Watanabe (Sept. 2018). "High-pressure protein crystal structure analysis of Escherichia coli dihydrofolate reductase complexed with folate and NADP+". In: *Acta Crystallographica Section D Structural Biology* 74.9, pp. 895–905. DOI: 10.1107/s2059798318009397. URL: http://dx.doi.org/10.1107/S2059798318009397.

Oliveira, A. Sofia F. et al. (July 2021). "Dynamical nonequilibrium molecular dynamics reveals the structural basis for allostery and signal propagation in biomolecular systems". en. In: *The European Physical Journal B* 94.7. DOI: 10.1140/epjb/s10051-021-00157-0. URL: http://dx.doi.org/10.1140/epjb/s10051-021-00157-0.

Parrinello, M. and A. Rahman (Dec. 1981). "Polymorphic transitions in single crystals: A new molecular dynamics method". en. In: *Journal of Applied Physics* 52.12, pp. 7182–7190. DOI: 10.1063/1.328693. URL: http://dx.doi.org/10.1063/1.328693.

Pettersen, Eric F. et al. (2004). "UCSF Chimera A visualization system for exploratory research and analysis". en. In: *Journal of Computational Chemistry* 25.13, pp. 1605–1612. DOI: 10.1002/jcc.20084. URL: http://dx.doi.org/10.1002/jcc.20084.

Ponder, Jay W. and David A. Case (2003). "Force Fields for Protein Simulations". In: *Protein Simulations*, pp. 27–85. DOI: 10.1016/s0065-3233(03)66002-x. URL: http://dx.doi.org/10.1016/s0065-3233(03)66002-x.

Reynolds, Kimberly A., Richard N. McLaughlin, and Rama Ranganathan (Dec. 2011). "Hot Spots for Allosteric Regulation on Protein Surfaces". en. In: *Cell* 147.7, pp. 1564–1575. DOI: 10.1016/j.cell.2011.10.049. URL: http://dx.doi.org/10.1016/j.cell.2011.10.049.

Royer William E., Jr. et al. (Dec. 1996). *Ordered water molecules as key allosteric mediators in a cooperative dimerichemoglobin*. en. DOI: 10.1073/pnas.93.25.14526. URL: http://dx.doi.org/10.1073/pnas.93.25.14526.

Sterpone, Fabio, Guillaume Stirnemann, and Damien Laage (Feb. 2012). *Magnitude and Molecular Origin of Water Slowdown Next to a Protein*. en. DOI: 10.1021/ja3007897. URL: http://dx.doi.org/10.1021/ja3007897.

Swain, J and L Gierasch (Feb. 2006). *The changing landscape of protein allostery*. en. DOI: 10.1016/j.sbi.2006.01.003. URL: http://dx.doi.org/10.1016/j.sbi.2006.01.003.

Swope, William C. et al. (Jan. 1982). "A computer simulation method for the calculation of equilibrium constants for the formation of physical clusters of molecules: Application to small water clusters". en. In: *The Journal of Chemical Physics* 76.1, pp. 637–649. DOI: 10.1063/1.442716. URL: http://dx.doi.org/10.1063/1.442716.

Togashi, Yuichi and Holger Flechsig (Dec. 2018). *Coarse-Grained Protein Dynamics Studies Using Elastic Network Models*. en. DOI: 10.3390/ijms19123899. URL: http://dx.doi.org/10.3390/ijms19123899.

Vanommeslaeghe, K. et al. (2010). "CHARMM general force field: A force field for drug-like molecules compatible with the CHARMM all-atom additive biological force fields". In: *Journal of Computational Chemistry* 31.4,

pp. 671–690. ISSN: 1096-987X. DOI: 10.1002/jcc.21367. URL: http://dx.doi.org/10.1002/jcc.21367.

Villaverde, Antonio (Oct. 2003). *Allosteric enzymes as biosensors for molecular diagnosis*. en. DOI: 10.1016/s0014-5793(03)01160-8. URL: http://dx.doi.org/10.1016/s0014-5793(03)01160-8.

Wodak, Shoshana J. et al. (Apr. 2019). *Allostery in Its Many Disguises: From Theory to Applications*. en. DOI: 10.1016/j.str.2019.01.003. URL: http://dx.doi.org/10.1016/j.str.2019.01.003.

A System preparation

The PDB structure 5Z6F (Nagae, Yamada, and Watanabe 2018) has been used. A mutation was present in the chain and it has been removed by using Chimera (Pettersen et al. 2004), with the rotamers tool.
The forcefield which has been used is AMBER99SB (Ponder and Case 2003), and therefore the initial topologies for the ligands have been generated by using ACPype.
AMBER has been chosen in spite of CHARMM (Brooks, Brooks III, and al. 2009) because there are problems with the generation of the forcefield of the ligand using CGenFF (Vanommeslaeghe et al. 2010), and there were not many parameters published online to be used for it.
Also in the case of AMBER, DHF does not behave quite well without proper ab-initio calculations to obtain the partial charges and atom-atom interactions within the molecule. For this reason we used the partial charges found in literature (Liu et al. 2013).
The system was solvated inside a cubic box, where 10205 water SPC/E molecules are added to the system, and to neutralize the charge, 15 sodium ions are added to the system.
The system then undergoes energy minimization, in order to start sampling the system from its lowest possible energy: steepest descent is used for these steps, and it continues until the maximum force in the system is less than $50\,\mathrm{kg\,mol^{-1}nm^{-1}}$.
The protein and the ligands are restrained to their initial positions, by applying a constant force in the 3 directions to all their heavy atoms.
An NVT simulation to equilibrate the system to a given temperature is performed, and the chosen thermostat is velocity rescaling by Bussi-Parrinello. The temperature is slowly increased from 1K to 300K to avoid large initial fluctuations in the ligands, which could compromise the correct structure of the complex itself.
Finally, an NPT simulation is performed, by using Berenden's algorithm.
The restraints which kept the complex still are then released and the system is then simulated for 50ns, using Parrinello-Rahman's barostat instead of Berendsen's.
Most of the values for neighbour searching or interaction approximations in all the simulations are left to the default values defined in GROMACS.

To analyze the system, the system trajectory is converted so that the system can go outside of the box (in order to compute distances in the correct way), and the rotational and translational movement of the system is removed, in order to avoid artefacts in the analyses.

B System preparation: water analysis

Trajectories have been generated, for both closed and open conformations of DHFR.
Both systems are equilibrated as described in previous section.
In this case we will store the positions of all particles at intervals of 10 femtoseconds, instead of the order of picoseconds.
This is necessary, in order to capture the time-scales related to water. The systems are then sampled for 10ns and the previously mentioned tool has been used to extract the water reorientational times.

C Data tables

C.1 HCRs

Connected residues:[4, 5, 13, 14, 15, 16, 27, 28, 29, 31, 33, 45, 46, 54, 94, 95, 96, 97, 98, 99, 100, 101, 102, 103, 104, 123, 124, 125]
Not connected residues: [68, 107]
HCRs obtained as described in the correlation section, with a distance threshold between Cα atoms of 4Å, and absolute value of the correlation larger than 0.06. p-value smaller than 2×10^{-5}.

C.2 Slow sites

Connected residues: [3, 22, 30, 45, 81, 90, 91, 92, 112, 121, 123, 127, 133]
Not connected residues: [85, 105, 108, 115]
Slow sites obtained as described in the water analysis section, with a distance threshold between Cα atoms of 4Å, and reorientational time threshold larger than 35ps. p-value around 0.03.